Journal of Telematics and Informatics (JTI)
Vol.6, No.3, September 2018, pp. 230~237
ISSN: 2303-3703

Sentiment Analysis of Indonesian Figure using Support Vector Machine

Suharyo Herwasto [1], Imam Much Ibnu Subroto [2], Badieah Assegaf [2]
[1)] Student of Informatics Engineering Department, Islamic University of Sultan Agung Semarang
[2)] Informatics Engineering Department, Islamic University of Sultan Agung Semarang
Email: suharyo.h@std.unissula.ac.id, imam@unissula.ac.id, badieah@unissula.ac.id

Abstract

On the political year 2018 will be mutually popping reverberated figures for Indonesian presidential candidate 2019. The figures recognition process generally are now using social media, so it would appear the opinions of social media users. Opinions that appeared not only contain positive and negative polarity, but also contain a sentence of subjective and objective. By using a machine learning algorithm, namely Support Vector Machine, made sentiment analysis. The results of the analysis of this sentiment more optimally use the kernel Linear with the F-Measure of Polarity 68%, 68%, 63%, and the F-Measure Subjectivity 73%, 77%, 75% for each figure Anies Baswedan, Joko Widodo, and Prabowo Subianto.

Keywords: *Sentiment analysis, support vector machines, figure candidate of Indonesian President 2019*

1. Introduction

Entering 2018, Indonesia entered the political year, where participants elections begin preparations for the democratic party presidential elections once every five years that will take place in 2019. Various pollsters have conducted surveys field to determine electability against a figure of the presidential candidate , The survey by the Indonesian Survey Circle[1] and Indo Barometer [2] placing 3 figures of the presidential candidates with the highest electability, the figure is Joko Widodo, Prabowo Subianto , and Anies Baswedan.

Three figures are a well-known figure. Many news preach the figures, as well as the people who talk about the three figures. Society at this time either supporting or not much discussion on social media. According to Sudibyo [3] and Kadarsih [4] social media serves as a provider of public opinion facility.

Opinions that appears from public opinion not only in the polarity (positive opinion or negative opinion), but the opinion is also shaped subjectivity opinions (opinions arises whether based on real circumstances or by opinion makers thought).

This research will use machine learning techniques, the method of Support Vector Machine (SVM) to analyze the opinion (sentiment) based on polarity and subjectivity.

The parts of this study are as follows: Section II describes a similar study has been done. Section III describes the machine learning technique for sentiment analysis. Section IV describes the experimental setup and evaluation. And in Section V is conclusions and recommendations of this study.

2. Related Work

Machine learning method is used to analyze the data of the Turkish state news on whether the author gives a criticism or support for political parties, politicians, or social issues. By comparing the methods Naïve Bayes, Support Vector Machine, Maximum Entrophy, and N-Gram Language Model. In the research process, not using stopword because stopword can provide an strong effects on sentiment. Total power 200 columnis positive data and 200 negative columnis data to create 3-fold cross validation, the results obtained accurately the method used between 65% to 77%[5].

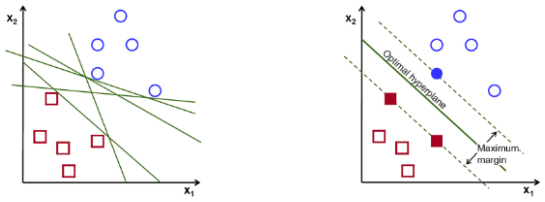
Ni Wayan Sumartini Saraswati conduct research to analyze sentiment using Naïve Bayes and Support Vector Machine to the English-speaking movie reviews and opinion data from Bali rubric Latest loaded in the Bali Post. Using data from 5000 for each positive and negative sentiment showed that accuracy results with Support Vector Machine is better than Naïve Bayes on opinions of Indonesian language and English[6].

In a study of online news classification using Support Vector Machine and K-Nearest Neigbour. By taking data of 500 news sites from detik.com where it will be distributed each 100 news to 5 categories accuracy results obtained for Support Vector Machine method is better than K-Nearest Neighbor (KNN) where the value of the test (accuracy, precision, recall, and F-Measure) reached more than 90% were using a linear or polynomial kernel. While KNN test values obtained between 60% to 81%[7].

Ni Wayan Sumartini Saraswati melakukan penelitian untuk menganalisa sentimen menggunakan metode Naïve Bayes dan Support Vector Machine terhadap review film yang berbahasa Inggris dan data opini dari rubrik Bali Terkini yang dimuat pada Bali Post. Dengan menggunakan data masing-masing 5000 untuk sentimen positif maupun negatif didapatkan hasil bahwa hasil akurasi dengan Support Vector Machine lebih baik dibandingkan Naïve Bayes pada opini berbahasa Indonesia maupun berbahasa Inggris [6].

3. Metode Machine Learning

Machine learning methods that will be used is a Support Vector Machine (SVM). The basic idea is to maximize the boundary hyperplane SVM. There are a number of options that may hyperplane for the set of data, but would have hyperplane with maximum margin[8].



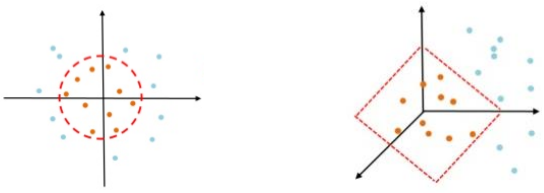
(a) (b)

Fig. 1. SVM Concept

In the picture there are several possible 1.a there are possibility for hyperplane to separate the data, but a good hyperplane is the hyperplane that has the maximum margin. Hyperplane with maximum margin shown in Figure 1.b.

SVM in divided SVM classification Linear and Nonlinear SVM. Linear SVM classification is the process known data pattern, the data pattern can be separated by a hyperplane linearly. While Nonlinear SVM classification is the process that the data pattern is not known. In the process of this sentiment analysis classification process will use Nonlinear SVM.

Pattern data that can not be separated linearly to be separated is used a kernel approach. Kernel in question is a function that maps the data features of the original dimension (low) to another feature higher dimension[8].



(a) (b)

Fig. 2 Pemetaan Dimensi

Kernel algorithms are shown as

$$\Phi: D^q \rightarrow D^r \quad (1)$$
$$x \rightarrow \Phi(x)$$

Φ kernel function used for mapping to a higher dimension. Here are some kernel functions in Table 1.

Table 1 Fungsi Kernel

Kernel Name	Fuction
Linear	$K(x,y) = x.y$
Polynomial	$K(x,y) = (x.y + c)^d$
Gaussian RBF	$K(x,y) = exp\left(\frac{-\| x-y \|^2}{2.\sigma^2}\right)$
Sigmoid	$K(x,y) = tanh(\sigma(x.y) + c)$

4. Experiment Setup and Evaluation

a. Experiment Setup

The data used is the data sentiment tweet in June 2018 with a portion of the data in Table 2 and Table 3.

Table 2 Dataset Sentiment Polarity

Figur Bakal Calon Presiden	Sentiment Polarity		Total	Data Training	Data Testing
	Positive	Negative			
Joko Widodo	120	120	240	176	64
Prabowo Subianto	120	120	240	176	64
Anies Baswedan	120	120	240	176	64

Table 3 Dataset Sentiment Subjectivity

Figur Bakal Calon Presiden	Sentiment Subjectivity		Total	Data Training	Data Testing
	Sub-jective	Ob-jective			
Joko Widodo	120	120	240	176	64
Prabowo Subianto	120	120	240	176	64
Anies Baswedan	120	120	240	176	64

The amount of data used in this study were 720 data where all the data is own sentiment polarity and sentiment subjectivity are balanced.

This tweet data is retrieved using the Twitter API with a free service.

The research model that will be used are as in Figure 3.

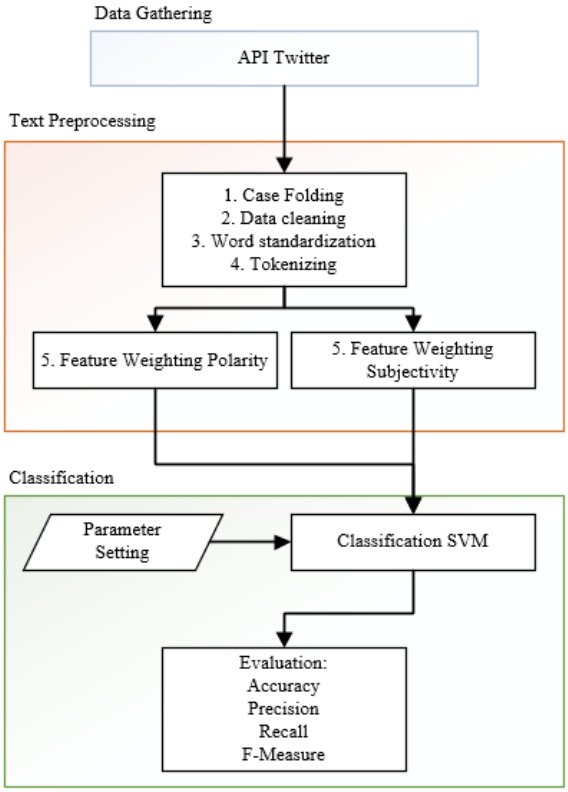


Fig. 3. Research Method

This research method is divided into three main sections, namely the data gathering, the data preprocessing and classification.

Data gathering tweet this research using the data in Table 2 and 3.

Text preprocessing consists of the folding case, data cleaning, word standardization, and tokenizing. Folsing case process is a process to change the code on a tweet from uppercase letters into lowercase. While the data cleaning stage is the process to remove words that do not contain such sentiments mention to an account that begins with the "@" and url links that participate in the content of tweets that begin with "http" or "https". Word standardization is the process of changing the word abbreviations into the normal form of the word. In the process of standardization needed a dictionary word in which there are a list of abbreviations and words normal. After word tokenizing process standardization is done, the process is based on sentiment aware tokenizing and negation marking. Tokenizing this way produces a value greater accuracy than tokenizing only by spaces or whitespace [9].

Because the sentiment to be analyzed is based on polarity and subjectivity, then for feature weighting process will be isolated as well. Feature weighting is done using TF-IDF Delta. Delta Mechanical TF-IDF is a development of the technique TF-IDF and the results of increased accuracy [10].

In the classification phase, the classification process for sentiment and subjectivity sentiment polarity done separately. This method is done for every figure, so that there will be 3 process research and figure studies done for the weighting and feature a separate classification for analyzing the polarity and subjectivity. After the classification results are known, the further process of evaluation. The evaluation process uses three aspects of assessment, namely accuracy, precision and recall.

b. Evaluation

To evaluate the results of sentiment classification, this study used an assessment of accuracy, precision, and recall.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (2)$$

$$Precision = \frac{tp}{tp + fp} \quad (3)$$

$$Recall = \frac{tp}{tp + fn} \quad (4)$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \quad (5)$$

5. Experiment Result

By using Linear and Gaussian RBF kernel, classification results are as shown in Table 4.

Table 4 Polarity Sentiment Kernel Rbf

Figure	Accuracy	Precision	Recall	F-Measure
Anies Baswedan	70	62	74	68
Joko Widodo	73	72	74	73
Prabowo Subianto	75	67	81	73

Table 5 Subjectivty Sentiment Kernel Rbf

Figure	Accuracy	Precision	Recall	F-Measure
Anies Baswedan	67	58	73	64
Joko Widodo	72	81	68	74
Prabowo Subianto	66	79	63	70

By using a Gaussian kernel RBF and parameters of the gamma form 10^{-3} setting up 10^{2} and C from 10^{-2} up to 10^{2} obtain best results is the value of gamma = 10^{-3} and the value of C = 10^{2} the test values in Table 4 and Table 5.

Table 6 Polarity Sentiment Kernel Linear

Figure	Accuracy	Precision	Recall	F-Measure
Anies Baswedan	75	56	86	68
Joko Widodo	69	66	70	68
Prabowo Subianto	69	52	81	63

Table 7 Subjectivity Sentiment Kernel Linear

Figur	Akurasi	Presisi	Recall	F- Measure
Anies Baswedan	73	70	77	73
Joko Widodo	75	84	71	77
Prabowo Subianto	71	88	66	75

While using the Linear kernel and parameter setting C 10^{-2} to 10^{2}, obtain best test value is the value of C = 10^{-2} the test values in Table 6 and Table 7.

In the classification of sentiment polarity, RBF kernel better than linear kernel, but in subjectivity sentiment classification, Linear kernel better than RBF kernel.

In Handbook of Natural Language Processing second edition [11], The sentence is a sentence explaining opinionated opinions explicitly or implicitly. In it can be subjective and objective sentence. Thus, in the process of sentiment analysis is indeed necessary to sentiment analysis of subjectivity.

A Positive-subjective means that the tweet is a tweet that supports a figure but the sentences in the tweet based on personal opinions. A Negative-subjective means that the tweet is a tweet that is anti or does not support a figure and the sentences in the tweet based on personal opinions. So that subjective tweets indicate that the person who made the tweet was an emotional person. A positive-objective is a tweet that support a figure and the sentences in a tweet containing a facts about the figure, and a negative-objective is a tweet that is anti or does not support a figure and the sentences in the tweet containing a facts about the figure. Tweets that contain objective sentences indicate that the Twitter user is a rational person.

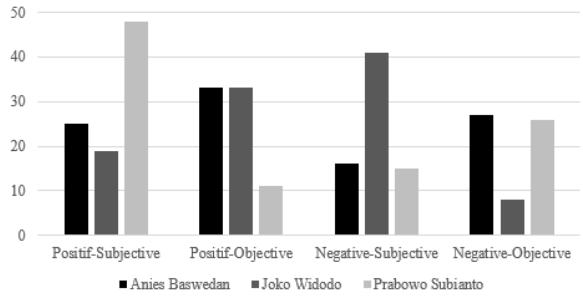


Fig. 4 Polarity Influenced by Subjectivity using Kernel RBF

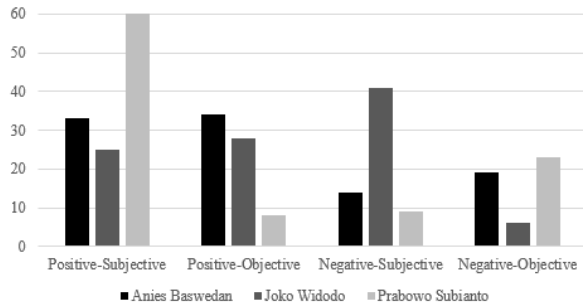


Fig. 5 Polarity Influenced by Subjectivity using Kernel Linear

In Fig. 4 and Fig.5, although it has different values but in general the level of influence of the kernel have the same value, the figure Anies Baswedan highest value is on the positive sentiment that is influenced by the objectivity, Joko Widodo figures appear much negative sentiment that effected by a subjectivity and figure Prabowo Subianto has a lot of positive sentiment is effected by subjectivity.

One of data visualization form from sentiment analysis that uses Twitter data is a sentiment map. The sentiment map is an illustration of the shape of the map of Indonesia which is divided into provinces and each province will be marked with colors. Provinces with a green color indicate that the province tends to be positively oriented towards a figure that is determined, provinces with a yellow color indicate the number of positive and negative sentiments are the same, red color indicates that the province tends to have a lot of negative sentiments, and the gray color indicates that in the province there is no tweet data. The form of the sentiment map looks like Fig 6.

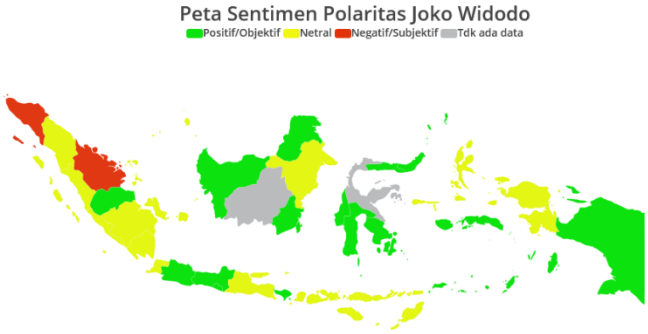


Fig. 6 The Sentiment Map

6. Conclusion

Classification of the test results showed that the average use linear kernel using a value of c = 10^{-2} better than using a gaussian kernel RBF.

For further research may use more data. With more data, machine learning can also learn more and can deliver results better accuracy. In addition besides the more data, use the Twitter API

to the data gathering is better to use a paid service because it can speed up the process of labeling sentiment data because the data obtained is not cut when the sentiment is more than 140 characters. In addition it can be used to other methods to be able to know which methods are good for sentiment analysis.

References

[1] TEMPO.CO. (2018). *Survei: Prabowo Peringkat Atas Penantang Jokowi di PIlpres 2019*. Available: https://nasional.tempo.co/read/1056788/survei-prabowo-peringkat-atas-penantang-jokowi-di-pilpres-2019

[2] Merdeka.com. (2018). *Selain Prabowo, Anies Baswedan dinilai saingan terberat Jokowi di Pilpres 2019*. Available: https://www.merdeka.com/politik/selain-prabowo-anies-baswedan-dinilai-saingan-terberat-jokowi-di-pilpres-2019.html

[3] A. Sudibyo, *Politik media dan pertarungan wacana*. LKiS, 2001.

[4] R. Kadarsih, "Demokrasi dalam Ruang Publik: Sebuah Pemikiran Ulang untuk Media Massa di Indonesia," *Jurnal Dakwah,* vol. 9, no. 1, pp. 1-12, 2008.

[5] M. Kaya, G. Fidan, and I. H. Toroslu, "Sentiment analysis of turkish political news," in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, 2012, pp. 174-180: IEEE Computer Society.

[6] N. W. S. Saraswati, "Naïve Bayes Classifier Dan Support Vector Machines Untuk Sentiment Analysis," *SESINDO 2013,* vol. 2013, 2013.

[7] S. N. Asiyah and K. Fithriasari, "Klasifikasi Berita Online Menggunakan Metode Support Vector Machine dan K-Nearest Neighbor," *Jurnal Sains dan Seni ITS,* vol. 5, no. 2, 2016.

[8] E. Prasetyo, *Data Mining - Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: PENERBIT ANDI, 2012.

[9] C. Potts. *Sentiment Symposium Tutorial: Tokenizing*. Available: http://sentiment.christopherpotts.net/tokenizing.html

[10] J. Martineau and T. Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis," *Icwsm,* vol. 9, p. 106, 2009.

[11] N. Indurkhya and F. J. Damerau, *Handbook of natural language processing*. CRC Press, 2010.